Université Lumière Lyon 2
UFR ICOM – Master Informatique
Master Informatique, 1ère année

Rapport de Stage

Analyse et modélisation prédictive des effectifs universitaires

Abdourahmane Timera

Stage effectué à la
Division des Études Statistiques (DES) – UCAD

Du 02/05/2025 au 31/06/2025

Année universitaire 2024–2025

Tuteur en entreprise :
Dr. Mountagua Lam
(DES – UCAD)

Encadrant pédagogique :
Ricco Rakotomalala
(Université Lumière Lyon 2 – ICOM)

Table des Matières

Table des Matières

1 Introduction Generale

L'université cheikh Anta Diop produit chaque année une quantité importante de données relatives aux étudiants. ces données sont essentielles pour la gestion et la planification des ressources académiques.l'analyse de ses donnes peuvent permettre a l'université de tirer des enseignements précieux sur les tendances d'inscription, les performances académiques et anticiper les besoins futures .

C'est dans ce contexte que j'effectue mon stage au sein de la Division des Études Statistiques (DES) de cette université, une structure charge de réalise des analyses statistiques et de fournir des informations pertinentes pour la prise de décision au sein de l'université.

Durant ce stage j'ai ete implique dans diverses taches allant de la compréhension des donnes a l'implementation d'un modèle de machine learning . Cette expérience m'a permis de développer mes compétences en analyse de données, en programmation et en visualisation des données tout en découvrant la problématique des universités africaines .

Ce Rapport présente en details les objectifs de mon stage, les technologies utilisées, les résultats obtenus ainsi que les competences acquises.il proposera quelques pistes de réflexion pour l'amélioration continue de l'analyse des données au sein de l'université.

2 Presentation de l'organisme d'accueil

2.1 Direction des Études et des Statistiques (DES)

la Direction des Études et des Statistiques (DES) a pour mission principale de fournir aux autorités de l'Université, aux établissements ainsi qu'aux différentes directions et services, des informations fiables et pertinentes sous forme d'études, d'analyses, de tableaux de bord et d'indicateurs. Ces outils permettent d'éclairer les prises de décisions stratégiques et opérationnelles au sein de l'Université.

La DES est responsable de la collecte, de la centralisation, de la supervision, de l'analyse et de la diffusion des données statistiques de l'UCAD. Elle veille à garantir la qualité, la cohérence et la fiabilité des informations produites, en harmonisant notamment le calcul des indicateurs et en validant, sous l'approbation de la commission de validation, toute production statistique avant communication publique.

En somme la DES facilite la prise de decision au sein de l'Université en fournissant des insights pertinents bases sur une analyse rigoureuse des données.

2.2 Mission de la DES

Plusieurs divisions existent au sein de la DES, chacune ayant des missions spécifiques :

- **Division planification suivi-Evaluation (DivPSE)** : Leur mission principales sont les suivantes :
  - Suivre l'orientation des programmes et projets .
  - Mettre en oeuvre le suivi de ces programmes et projets.
  - Produire des rapports de performances des reformes annuelles
- **Division des Études et suivi de la performance** : ils sont chargés de :

  - Calculer et mettre a jour les indicateurs de performance
  - Réaliser des Enquêtes de satisfaction avec la communication .

- **Division des Études Statistiques (DES)** : c'est la division dans laquelle j'ai effectué mon stage. Elle est chargée de :
  - Produire la section de l'annuaire statistique de l'UCAD.
  - Collecter, traiter et analyser les données statistiques
  - Soutenir le development des systèmes de collectes de données .

voici un organigramme de la DES et ses divisions :

Responsabilités par Division au sein de la D.E.S

Caractéristique	Division des Études et Suivi de la Performance	Division des Statistiques	Division de la Planification et Suivi-Évaluation
Mesure de la Performance	Assurer la cohérence et la qualité	Produire la section de l'annuaire statistique	Suivre l'orientation des programmes et projets
Gestion des Données	Calculer et mettre à jour les indicateurs de performance	Soutenir les opérations de saisie des données statistiques	Mettre en œuvre le suivi des programmes et projets
Campagnes Statistiques	Participer aux campagnes de données statistiques	Participer aux campagnes de données statistiques	Produire des rapports de performance des réformes annuelles
Enquêtes de Satisfaction	Réaliser des enquêtes de satisfaction avec la Communication	Collecter, traiter et analyser les données statistiques	Promouvoir la coopération avec les partenaires techniques et financiers
Annuaire Statistique	Participer à la diffusion de l'annuaire statistique	Produire l'annuaire statistique	Soutenir la gouvernance, la formation, les études de recherche
Stockage des Données	Interface avec la DISI	Stocker, classer et archiver les fichiers numériques	Aider au développement des outils de planification
Systèmes de Collecte de Données	Soutenir les opérations de saisie des données statistiques	Soutenir le développement des systèmes de collecte de données	Suivre l'exécution du plan stratégique et évaluer les résultats
Reporting	Participer à la diffusion des rapports de performance	Participer à la diffusion des rapports de performance	Contribuer aux études prospectives sur les politiques publiques

Figure 1: Organigramme de la Division des Études Statistiques (DES)

3 Mission de stage

3.1 Contexte du stage

Les universités Africaines plus précisément l'université cheikh Anta Diop de Dakar font face a une multitude de problèmes une croissance rapide des effectifs étudiants, des inscription de plus en plus tardives, des taux d'abandon élevés et une gestion des ressources académiques souvent inefficace liées a une manque de ressources financières et humaines. Dans ce contexte, l'analyse des données devient un outil essentiel pour comprendre les tendances d'inscription, les performances académiques et anticiper les besoins futurs. C'est dans ce cadre que j'ai effectué mon stage au sein de la Division des Études Statistiques (DES) pour essayer a travers l'analyse des données d'apporter des solutions aux problèmes rencontrés par l'université.

3.2 Objectifs du stage

L'objectif principal de mon stage était de contribuer à l'analyse des données relatives aux étudiants de l'université. Plus précisément, il s'agissait de :

- collecter,nettoyer et structurer les données historiques sur les effectifs des étudiants.
- Réaliser des analyses descriptives pour identifier les tendances les variations saisonnières et les anomalies dans les données.
- Développer un modèle de machine learning facilitant ainsi la planification stratégique et la prise de décision au sein de l'université.
- Présenter les analyses et recommandations de manière claire décideurs de l'université

3.3 Difficultés rencontrées

Au cours de mon stage, j'ai rencontré plusieurs difficultés :

- **Qualité des données** : Absence de texte pour expliquer explicitement les colonnes dans la base de données, ce qui a rendu la compréhension des données plus difficile.
- **Accès aux données** : Difficultés à accéder à certaines données historiques en raison de la structure des bases de données et des restrictions d'accès.
- **Complexité des analyses** : Avec plusieurs faculté , départements et ecoles les données étaient volumineuses et complexes, nécessitant des techniques avancées de nettoyage et d'analyse.
- **Intégration des données** : Fusionner les données provenant de différentes sources a été un défi en raison de la diversité des formats et des structures.

4 Données et Technologies Utilisés

4.1 Sources et Descriptions des données

4.1.1 Sources des données

Les sources de données proviennent d'un entrepôt de données appelé RADIS qui regroupe toutes les données de l'université, notamment les données relatives aux étudiants, aux enseignants et aux ressources académiques. Pour accéder aux données, je me suis connecté avec les identifiants fournis par la division. Une fois connecté, j'ai effectué des requêtes via l'interface en ciblant les tables pertinentes. Les informations nécessaires ont été choisies en amont par la Division des Études Statistiques (DES) pour répondre aux besoins d'analyse. Enfin, les données ont été extraites au format CSV pour faciliter leur manipulation et analyse ultérieure.

4.1.2 Description des données

Les données sont des données étudiants de plusieurs années couvrant une période allant de 2001 à 2025. Les données extraites comprennent plusieurs informations sur les étudiants, notamment :

- **Identifiant étudiant** : Un identifiant unique pour chaque étudiant.
- **INE Étudiant** : Un identifiant national étudiant (INE) unique pour chaque étudiant, utilisé pour l'identification officielle.
- **Nom et prénom, sexe** : Les noms, prénoms et le sexe des étudiants.
- **Date de naissance** : La date de naissance des étudiants.
- **Lieu et région de naissance** : Le lieu et la région de naissance de l'étudiant.
- **Faculté** : La faculté à laquelle l'étudiant est inscrit.
- **Département** : Le département spécifique au sein de la faculté.
- **Année d'inscription** : L'année d'inscription de l'étudiant.
- **Niveau d'étude** : Le niveau d'étude de l'étudiant : première année, deuxième année, etc.
- **Système inscrit** : Le système d'inscription de l'étudiant (par exemple LMD ou classique).

  On peut aussi trouver des informations sur les performances académiques antérieures de l'étudiant, notamment celles relatives au baccalauréat, telles que :

- **Année du baccalauréat** : L'année où l'étudiant a obtenu son baccalauréat.
- **Mention au baccalauréat** : La mention obtenue par l'étudiant au baccalauréat.

- **Série du baccalauréat** : La série du baccalauréat obtenue par l'étudiant.

  D'autres colonnes existent aussi dans la base de données mais ne sont pas pertinentes pour l'analyse que nous avons effectuée.

  Les données "resultats" ont été aussi extraites de l'entrepôt RADIS et comprennent des informations sur les notes. Ces deux bases ont presque les mêmes colonnes, excepté les colonnes suivantes :

- **Crédit** : Le nombre de crédits obtenus par l'étudiant.
- **Session** : La session d'examen (première ou deuxième session).
- **Mention** : La mention obtenue par l'étudiant pour la session d'examen.
- **Moyenne annuelle** : La moyenne annuelle de l'étudiant pour l'année académique.
- **Résultats** : Le résultat après délibération (admis, ajourné, etc.).

> **Remarque**
>
> Les résultats étudiés s'étendent sur la période de 2011 à 2024.

4.2 Technologies Utilisées

Pour mener à bien les missions de mon stage, j'ai utilisé plusieurs technologies et outils qui m'ont permis de traiter, analyser et visualiser les données efficacement. Voici un aperçu des principales technologies utilisées :

4.2.1 Python et ses bibliothèques

Python est un langage de programmation puissant utilisé pour le traitement de données, l'analyse et la visualisation. Nous avons choisi Python pour sa flexibilité, sa simplicité dans le domaine de la science des données. Il offre une large gamme de bibliothèques adaptées à l'analyse de données, notamment :

- **Pandas** : Pour la manipulation et l'analyse des données, notamment pour le nettoyage, la transformation et l'agrégation des données.
- **NumPy** : Pour les opérations mathématiques et statistiques sur les tableaux de données.
- **Matplotlib** et **Seaborn** : Pour la visualisation des données, permettant de créer des graphiques et des diagrammes pour mieux comprendre les tendances et les distributions.
- **Scikit-learn** : Pour le machine learning, utilisé pour développer des modèles prédictifs basés sur les données historiques.

4.2.2 Visual Studio Code

Visual Studio Code, appelé plus souvent VSCode, est un éditeur de texte très connu et largement utilisé dans la communauté des développeurs. Il offre un environnement facile à manipuler et personnalisable pour le développement de projets Python. J'ai fait le choix de VSCode pour sa simplicité, sa rapidité et ses nombreuses extensions qui facilitent le développement et la gestion de projets Python.

4.2.3 Power BI

Power BI est un outil de visualisation de données développé par Microsoft. Il permet de créer des rapports interactifs et des tableaux de bord à partir de diverses sources de données. C'est l'outil de visualisation le plus utilisé par la communauté des analystes de données. Son interface facile à utiliser et sa rapidité de création de visualisations m'ont permis de créer des graphiques et des tableaux de bord interactifs pour présenter les résultats de l'analyse des données de manière claire et concise.

4.2.4 Microsoft Excel

Excel est un tableur de calcul utilisé pour la manipulation et l'analyse de données. Il est largement utilisé pour des tâches simples de nettoyage, d'agrégation et de visualisation des données. J'ai utilisé Excel pour un petit aperçu des données et faire les premiers filtres pour mieux comprendre les données avant de les importer dans Python pour une analyse plus approfondie. Il m'a permis aussi de faire des filtres et des tableaux croisés dynamiques pour détecter les tendances et les anomalies dans les données.

4.2.5 Autres outils

D'autres outils et technologies ont été utilisés pour faciliter le travail, notamment :

- **Git** : Pour sauvegarder mon code en ligne et collaborer avec un autre développeur, permettant de suivre les modifications et de gérer les versions du code.
- **Jupyter Notebook** : Pour le développement interactif et la documentation du code Python, permettant de combiner le code, les visualisations et les explications dans un même document.

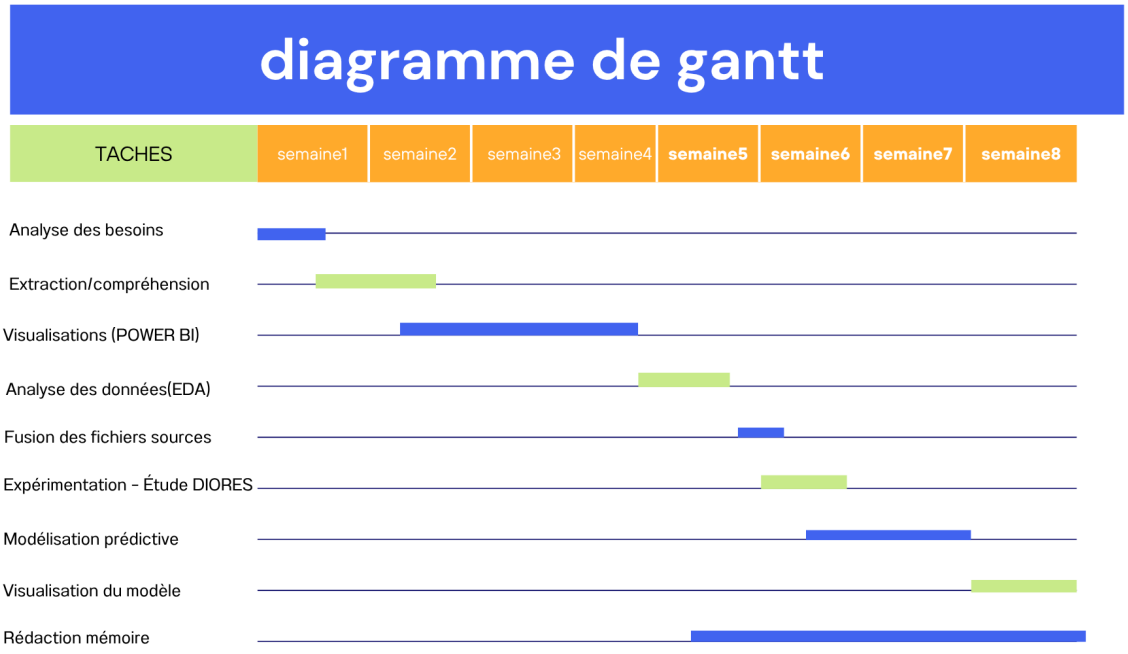
- **Slack** : Pour la communication et la collaboration avec l'équipe de la Division des Études Statistiques (DES), facilitant le partage d'informations et la coordination des tâches.

5 Travaux réalisés et résultats obtenus

5.1 Élaboration du diagramme de Gantt

Un diagramme de Gantt a été élaboré pour planifier et suivre les différentes étapes du projet. Voici un aperçu des principales étapes du projet :

- **Analyse des besoins** : Compréhension des objectifs du projet et des données disponibles.
- **Collecte des données** : Récupération des données nécessaires à l'analyse.
- **Nettoyage et préparation des données** : Traitement des données pour les rendre exploitables.
- **Analyse exploratoire des données (EDA)** : Exploration des données pour identifier les tendances et les anomalies.
- **Modélisation et analyse prédictive** : Développement de modèles pour prédire les résultats futurs.
- **Visualisation des données** : Création de graphiques et de tableaux de bord pour présenter les résultats.
- **Rédaction du rapport final** : Compilation des résultats et rédaction du rapport de stage. voici un aperçu du diagramme de Gantt :



5.2 Extraction et compréhension des données

Les données utilisées pour mon stage ont été extraites d'un entrepôt de données de l'Université appelé RADIS. Une fois extraites, les données ont été ouvertes sur plusieurs outils comme Excel et Power BI pour une première compréhension. Ensuite, elles ont été importées dans Python pour avoir un premier aperçu des données et faire les premiers filtres.

Pour bien comprendre les données, mon encadrant m'a demandé de faire plusieurs tâches comme :

- **Se rendre à la DISI** : Mon encadrant m'a demandé de me rendre à la DISI (Direction de l'informatique et des systèmes d'information) pour comprendre comment les données sont stockées, ce que signifie chaque colonne, et enfin prendre conscience des données disponibles pour des analyses futures.
- **Détection d'anomalies** : Une fois toutes les colonnes des données comprises, j'ai créé un fichier pour détecter les anomalies et incohérences dans les données. J'ai utilisé des outils comme Excel pour des filtres et des tableaux croisés dynamiques, ainsi que Python pour des analyses plus approfondies.

  Cette fonction prend en entrée un fichier de données CSV et retourne un fichier texte avec toutes les incohérences. Ce fichier, qui contient ces anomalies et incohérences, a été transmis à mon encadrant et à la DISI pour qu'ils puissent corriger les données.

  Cette tâche m'a permis de mieux comprendre les données et de m'assurer de leur qualité avant de procéder à l'analyse. J'ai aussi appris comment les données sont stockées et comment elles peuvent être utilisées pour des analyses futures. La difficulté de cette tâche était de comprendre les données et de détecter les anomalies, car les données étaient volumineuses et complexes. Cependant, j'ai réussi à surmonter cette difficulté en utilisant des outils comme Excel et Python pour filtrer et analyser les données.
- **Transformation des données** : Conversion des types de données, normalisation et agrégation des données pour les rendre exploitables.

5.3 Visualisation des données avec Power BI

Power BI a été utilisé pour créer des visualisations interactives des données. J'ai travaillé avec mon collègue qui avait déjà commencé à faire des visualisations sur Power BI. J'ai donc continue a travailler avec lui et on a réussi a faire plusieurs tableau de bord On a créer plusieurs graphiques et tableaux de bord pour présenter les résultats de l'analyse des données. Voici quelques exemples de visualisations que on a créées :

- **Tableaux de bord des étudiants** : On a créer un Tableau de bord des étudiants avec plusieurs graphiques et filtres pour explorer les données des étudiants par exemple pour chaque année le nombre d'étudiants inscrit homme et femme, le nombre d'étudiants pour chaque système LMD ou classique et pour chaque niveau (L1, L2 , 1A, 2A .....) , l'evolution des effectifs des étudiants au fil des années , le nombre d'étudiants inscrits pour chaque faculté , ecole , Institut et ecole doctorale. Voici quelques insights que nous avons pu obtenir à partir de ces visualisations :

  - le nombre d'étudiants inscrits a augmenté au fil des années, avec une tendance à la hausse significative. on compte plus de 80000 étudiants inscrits pour l'année académique 2023-2024.

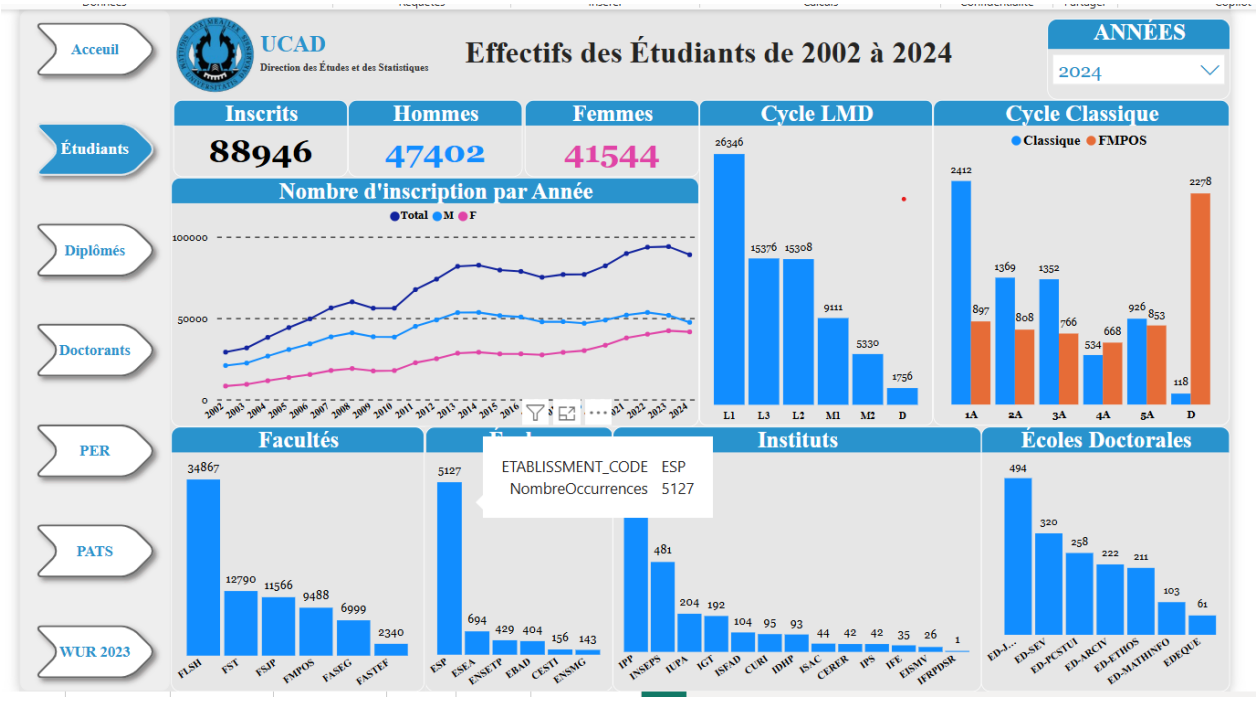
- 50% des étudiants sont inscrits a la faculté des lettres et sciences humaines (FLSH) et la faculté des sciences et techniques (FST)

- La majorité des étudiants sont inscrits en L1, suivis de L2 et L3.

- l'ecole avec le plus d'étudiants est L'ESP avec plus de 5000 inscrits pour l'année académique 2023-2024.

- la reparation d'hommes et femmes est presque égale, avec une légère majorité d'hommes 47000 hommes contre 41000 femmes pour l'année académique 2023-2024.

Voici un aperçu du tableau de bord des étudiants :



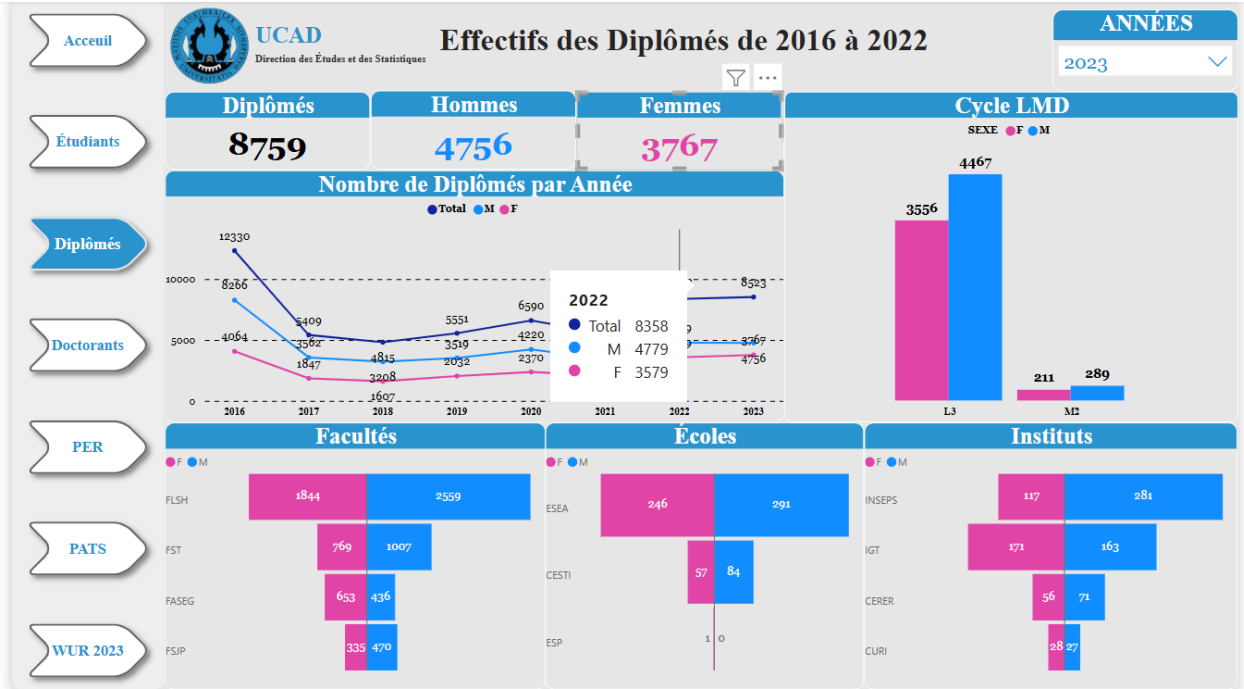
- **Tableaux de bord des diplômes** : Dans cette sections plusieurs graphiques ont été créés pour explorer les données des diplômes obtenus par les étudiants. On a pu créer des graphiques pour visualiser le nombre de diplômes obtenus par année, par faculté, par niveau d'étude . Voici quelques insights que nous avons pu obtenir à partir de ces visualisations :

  - le nombre de diplômés a diminuer entre 2016 et 2017 et est devenu constante jusqu'a 2021, avec une moyenne de 5000 diplômes (license et Master) par an et une légère augmentation en 2022 et 2023.

  - La majorité des diplômes sont issus de la la FLSH ce qui normalement est dû au nombre élevé d'étudiants inscrits dans cette faculté Cependant on remarque parmi ces diplômes, la majorité sont des licences et que ils sont constitue de 60% de femmes et 40% d'hommes.

  - Par contre avec l'ensemble des diplômes obtenus, on remarque que la majorité sont des hommes, avec 53% d'hommes et 47% de femmes.

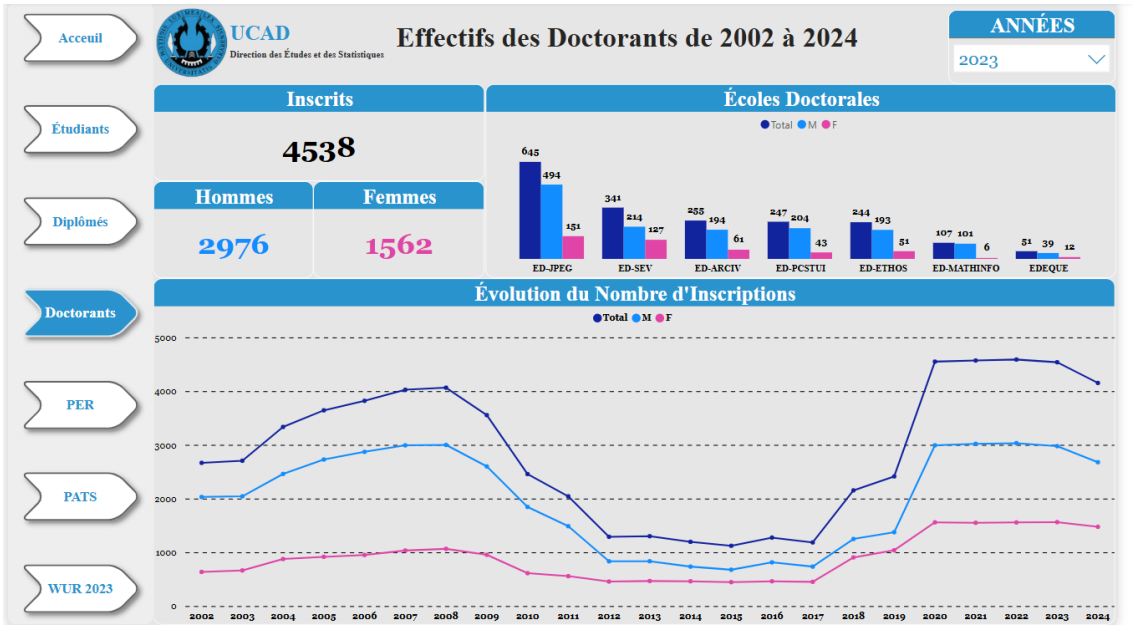
  Voici un aperçu du tableau de bord des diplômes :



- **Tableau de bord des doctorants** : Dans cette section, on a créé des graphiques pour explorer les données des doctorants de l'université. On a pu créer des graphiques pour visualiser le nombre de doctorants par année, par école doctorale, et par genre. Voici quelques insights que nous avons pu obtenir à partir de ces visualisations :

  - Le nombre de doctorants a augmenté au fil des années jusqu'en 2008, ensuite il a diminué jusqu'en 2012, puis il est resté constant jusqu'en 2017 et continue a augmenter jusqu'en 2023. On compte plus de 4500 doctorants avec 2900 hommes et 1562 femmes pour l'année académique 2023.

  - la majorité des doctorants sont inscrits a l'école doctorale JPEG avec plus de 600 inscription 400 hommes et 200 femmes pour l'année académique 2023-2024.



D'autres visualisations ont été créées pour explorer les données des étudiants, des diplômes et des doctorants. Ces visualisations ont permis de mieux comprendre les tendances et les anomalies dans les données, et de présenter les résultats de manière claire et concise.

5.4 Analyse exploratoire des données (EDA)

L'analyse exploratoire des données a ete réalise a l'aide de Python et de ses bibliothèques. Elle a permis de comprendre encore plus les les données et de détecter les tendances, les variations saisonnières. L'objectif de l'EDA est tout d'abord de mieux comprendre la structure des données , de detecter les anomalies et les incohérences deja détectées dans la section précédente , de réaliser des statistiques descriptives et de visualiser les données pour mieux comprendre les tendances et les variations saisonnières et afin de preparer les données pour la modélisation et l'analyse prédictive.

Pour faire cela j'ai tout d'abord écrit un petit programme en python qui permet de verifier si tous mes fichiers ont les memes colonnes et de les fusionner en un seul fichier

```
✅Les colonnes des fichiers sont identiques  au fichier de référence(fichier 2001-2002).
```

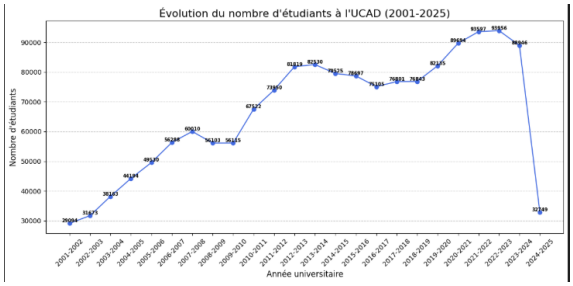
Une fois cette étape terminée, j'ai pu concaténer les fichiers en un seul appelé `df_total`. Après ces étapes, j'ai pu afficher les 5 premières lignes de mon fichier pour avoir un premier aperçu des données.
J'ai aussi utilisé les fonctions `info()` et `describe()` pour obtenir des informations sur les colonnes et des statistiques descriptives du fichier.

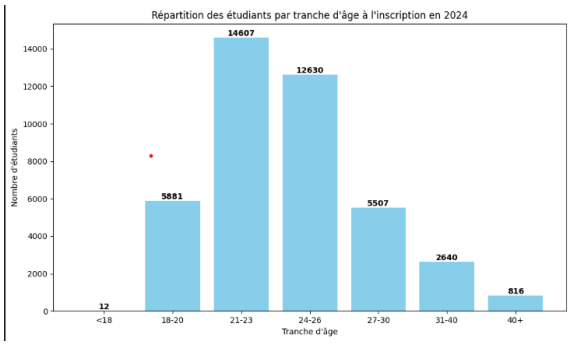
```
<class 'pandas.core.frame.DataFrame'>
Index: 88946 entries, 1473344 to 1562289
Data columns (total 33 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   NUMERO                   88946 non-null  object
 1   NUMERO_TABLE             51669 non-null  float64
 2   INE                      75953 non-null  object
 3   NUMERO_IDENTITIFICATION  88946 non-null  object
 4   NOM                      88946 non-null  object
 5   PRENOM                   88946 non-null  object
 6   DATE_DE_NAISSANCE        88946 non-null  object
 7   LIEU_DE_NAISSANCE        88946 non-null  object
 8   MAIL_INSTITUTIONNEL      88881 non-null  object
 9   SEXE                     88946 non-null  object
```

J'ai ensuite utilisé la fonction `isnull().sum()` pour vérifier s'il y a des valeurs manquantes dans le fichier. J'ai trouvé qu'il y avait des valeurs manquantes dans certaines colonnes, mais pas dans toutes les colonnes. voici les colonnes manquantes :

```
Tableau récapitulatif des valeurs manquantes :
+--------------------+---------------------+---------------+
| Colonnes           |   Valeurs manquantes |   Pourcentage |
|--------------------+---------------------+---------------|
| COHORTE            |                75789 |    85.2079    |
| NUMERO_TABLE       |                37277 |    41.9097    |
| NIVEAU LMD         |                15719 |    17.6725    |
| INE                |                12993 |    14.6077    |
| REGION_DE_NAISSANCE |                4602 |     5.17393   |
| SERIE_BACC         |                1265 |     1.42221   |
| MAIL_INSTITUTIONNEL |                  65 |     0.073078  |
| ANNEE_BACC         |                  20 |     0.0224856 |
| MENTION_BACC       |                  20 |     0.0224856 |
+--------------------+---------------------+---------------+
```

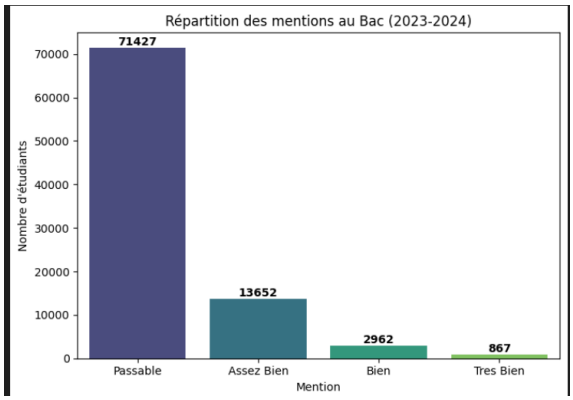
Apres avoir verifier les valeurs manquantes j'ai créer une autre cellule que j'ai renommé "données démographiques" pour faire une analyse des données démographiques des étudiants. j'ai pu observer l'evolution des effectifs des étudiants au fil des années, la répartition des étudiants par tranche d'age pour l'année 2024

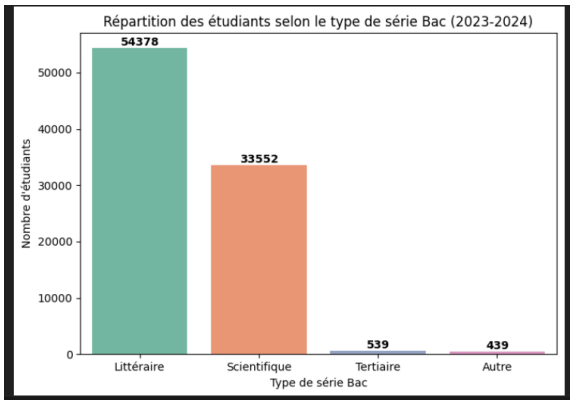




Conclusion : J'ai remarqué que la majorité des étudiants sont âgés de 18 à 25 ans, ce qui est normal car la plupart des étudiants commencent leurs études supérieures à cet âge. Cependant, il y a aussi une proportion importante d'étudiants âgés de 26 à 30 ans, ce qui peut être dû à des étudiants qui ont repris leurs études après une pause ou qui ont changé de filière. Concernant les inscriptions, j'ai pu observer l'évolution des inscriptions au fil des années. Cependant le nombre d'inscrits est tres basse en 2025 ce qui normale l'année 2024 n'est pas encore terminée.

Une fois cette etape termine une autre cellule a été crée pour voir l'historique académique des étudiants. Les différentes series et mentions ont été analysées pour comprendre la répartition des étudiants par série et par mention. J'ai utilisé la fonction `value_counts()` pour compter le nombre d'étudiants dans chaque série et chaque mention voici les résultats obtenus :

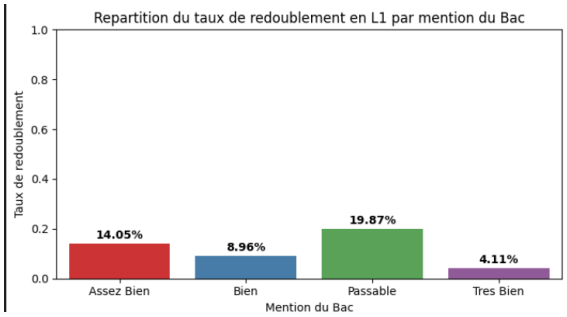


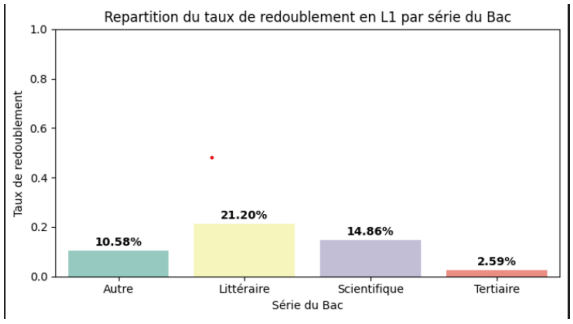


Conclusion: J'ai remarqué que la majorité des étudiants ont obtenu la mention "Passable", suivie de "Assez bien" et "Bien". la majorité des étudiants suivent un parcours littéraire(L2 , L1, A , AR etc..), suivi des parcours scientifiques(s1 , S2 , D , C ..) et Tertiaire

j'ai aussi observer le taux de redoublement en première année (l1 ou 1A)pour l'année 2023 est d'environ 19% ce qui est normal pour une université de cette taille.

j'ai aussi créer une cellule pour voir la proportion des ces redoublants par mention et par série du bac voici les résultats obtenus :





Conclusion :

- effet de la mention du Bac : Les résultats montrent que le taux de redoublement en L1 est d'autant plus élevé que la mention au Bac est faible. Ainsi, les étudiants ayant obtenu la mention "Passable" présentent un taux de redoublement de 19,87%, contre seulement 4,11% pour ceux ayant obtenu la mention "Très Bien". De même, les étudiants avec la mention "Assez Bien" connaissent un taux de redoublement relativement élevé 14,05%, tandis que ceux avec "Bien" affichent un taux plus faible 8,96%. Cela montre que la mention du Bac est un indicateur significatif de la réussite en L1, les étudiants ayant de meilleures mentions au Bac étant moins susceptibles de redoubler.

- Effet de la série du Bac : L'étude du taux de redoublement selon la série du Bac met en évidence une différence notable entre les profils:

Les étudiants issus des séries littéraires affichent le taux de redoublement le plus élevé 21,20%, suivis par ceux des séries scientifiques 14,86% et autres 10,58%.

Les étudiants issus des séries tertiaires enregistrent le taux le plus bas 2,59%. Cela suggère que les étudiants des séries littéraires et scientifiques rencontrent plus de difficultés en L1, tandis que ceux des séries tertiaires semblent mieux préparés pour cette étape universitaire.

Pour verifier si ces résultats sont significatifs, j'ai utilisé le test du Chi-2. J'ai utilisé la fonction `chi2_contingency()` de la bibliothèque `scipy.stats` pour effectuer le test du Chi-2. et j'ai constate que la p-value est inférieure à 0.05, ce qui signifie que les résultats sont significatifs. J'ai donc pu conclure que la mention du Bac et la série du Bac ont un effet significatif sur le taux de redoublement en L1.

Une autre analyse appelle suivi de cohorte a ete faite. Le but est de suivre une cohorte d'étudiants sur plusieurs années pour observer leur parcours académique. J'ai créer une fonction qui prend en entrée le jeux de donnée , l'année de bac , la série du bac et l'année universitaire et qui retourne le nombre d'étudiants qui on redouble pour chaque année universitaire , les diplômes apres 3 ans 4 ans et 5 ans et le taux d'abandon pour chaque année universitaire. Pour une premier version voici les résultats obtenus pour la cohorte de 2019 en série littéraire

```
suivi_cohorte(resultats, 2018, 'L2', '2018-2019')
```

```
**Synthèse de la cohorte** :
        Année            Niveau  Passe  Redouble  Abandon  Diplome
0   2018-2019                L1   2403      3404      169      0.0
1   2019-2020  L1 (redoublants)   1027         3     2374      0.0
2   2019-2020                L2   1464       873       66      0.0
3   2020-2021  L1 (redoublants)      0         0        3      0.0
4   2020-2021                L2    420       546       61      0.0
5   2020-2021  L2 (redoublants)    387        47      439      0.0
6   2020-2021                L3    220       497       32    715.0
7   2021-2022  L2 (redoublants)    130        11      452      0.0
8   2021-2022                L3     95       422       48    242.0
9   2021-2022  L3 (redoublants)     39        25      268    165.0
10  2022-2023  L2 (redoublants)      0         0       11      0.0
11  2022-2023                L3      4        79       12     35.0
12  2022-2023  L3 (redoublants)     12        18      251    166.0
13  2023-2024  L3 (redoublants)      0         0       95      2.0

Cohorte 2018 L2 (2018-2019) : 5976 étudiants inscrits en L1

Nombre de diplômés en 3 ans : 715
Nombre de diplômés en 4 ans : 407
Nombre de diplômés en 5 ans : 201
Nombre de diplômés en 6 ans : 2
```

Ici, on remarque que sur 5976 inscrits en L1 en 2018-2019, titulaires du baccalauréat série L2 obtenu en 2018, 1325 étudiants ont obtenu leur licence, soit environ 22,2% de la cohorte initiale. Parmi eux :

- 715 étudiants, soit 53.3% des diplomes, ont obtenu leur licence en 3 ans,
- 407 étudiants, soit 30.7%, en 4 ans,
- 201 étudiants, soit 15.16%, en 5 ans,
- enfin, 2 étudiants, soit environ 0,15%, en 6 ans.

On remarque aussi que 77,8% de cette cette cohorte n'obtienne pas la license (changement d'ecole exclusion voyage , etc...).

Une deuxième version de cette fonction a été écrite pour faire une analyse plus approfondie des résultats. cette fois si la fonction va prendre en paramètre la faculté et le département de l'étudiant en plus.

```
suivi_cohorte_versionII(resultats, 2018, 'L2', '2018-2019',
                        faculte='FSJP', departement='Droit Privé')
```

```
**Synthèse de la cohorte** :
        Année            Niveau  Passe  Redouble  Abandon  Diplome
0   2018-2019                L1    746      1525        0      0.0
1   2019-2020  L1 (redoublants)    406         1     1118      0.0
2   2019-2020                L2    268       469        9      0.0
3   2020-2021  L1 (redoublants)      0         0        1      0.0
4   2020-2021                L2    107       247       52      0.0
5   2020-2021  L2 (redoublants)    239        10      220      0.0
6   2020-2021                L3     39        99       10    120.0
7   2021-2022  L2 (redoublants)     47         4      206      0.0
8   2021-2022                L3     32       174       25    115.0
9   2021-2022  L3 (redoublants)     14         5       44     36.0
10  2022-2023  L2 (redoublants)      0         0        4      0.0
11  2022-2023                L3      0        30        2     15.0
12  2022-2023  L3 (redoublants)      1         0      113     65.0
13  2023-2024  L3 (redoublants)      0         0       30      0.0

Cohorte 2018 L2 (2018-2019) : 2271 étudiants inscrits en L1 | Faculté=FSJP | Département=Droit Privé

Nombre de diplômés en 3 ans : 120
Nombre de diplômés en 4 ans : 151
Nombre de diplômés en 5 ans : 80
Nombre de diplômés en 6 ans : 0
```

Conclusion : On remarque que de manière générale le taux de redoublement est relativement élevé. La plupart des étudiants redoublent en L1, suivis de L2 et L3 , et aussi beaucoup d'étudiants abandonnent leurs études en L1. A noter aussi que une grande partie des étudiants n'obtiennent pas leur licence..Cela peut être dû à plusieurs facteurs, tels que le manque de préparation, les difficultés d'adaptation à l'université , des problèmes personnels ou peut être le système d'orientation qui n'est pas adapté aux étudiants.

5.5 Fusion des données

Je dispose de plusieurs fichiers qui contiennent des informations sur les étudiants c'est a dire leur nom, prénom, date de naissance, numéro d'étudiant, etc. et leur parcours académiques et un autres fichier contenant les memes données avec les résultats universitaires en plus . J'ai donc decide de fusionner les deux fichiers pour avoir un seul fichier contenant toutes les informations sur les étudiants. Pour cela j'ai d'abord sélectionner les colonnes en commun entre les deux fichiers et j'ai utilisé la fonction `merge()` de la bibliothèque `pandas` pour fusionner les deux fichiers. Pour les fusionner j'ai utilise la colonne "Numéro" et "Année Université". j'ai utilise le paramètre "inner" pour ne garder que les lignes qui ont des valeurs communes dans les deux fichiers. Une fois la fusion terminée, j'ai vérifié si le nombre de lignes du nouveau fichier était égal à la somme des lignes des deux fichiers d'origine. Apres avoir vérifié que la fusion s'est bien déroulée, j'ai enregistré le nouveau fichier dans un fichier csv appelle `base_finale` .

```python
df_final = pd.merge(df_inscrit, #
                    df_resultat, #
                    on=['NUMERO', 'ANNEE UNIVERSITAIRE'], #
                    how='inner'
                    )
print("✅ Fusion des DataFrames df_inscrit et df_resultat réussie.")
```

Python

```
✅ Fusion des DataFrames df_inscrit et df_resultat réussie.
```

5.6 Etude du système DIORES

DIORES est un système intelligent d'orientation et de reorientation dans l'enseignement supérieur au Sénégal réalise par des chercheurs et des étudiants de l'Université Cheikh Anta Diop de Dakar. En effet le système actuelle Campusen n'oriente pas efficacement les étudiants dans l'enseignement supérieur. Il ne prend pas en compte des critères comme le profil , les réalités socio-économiques. Il ne propose que les recommandations base sur les données. Ce système n'est pas adapte aux besoins des étudiants car on constate un taux d'échec tres élevé dans l'enseignement supérieur. C'est dans ce contexte que l'etude vise a développer un système d'orientation base sur L'IA pour améliorer pour permettre aux étudiants d'être mieux oriente et réoriente dans l'enseignement supérieur. Les chercheurs analysent l'écart le classement d'un étudiant de Capusen et sa réussite réelle dans l'enseignement supérieur. Les chercheurs analysent l'écart entre le classement d'un étudiant par Campusen et sa réussite réelle dans l'enseignement supérieur, dans le but de proposer un modèle plus fiable. Pour cela, ils ont utilisé des algorithmes d'apprentissage automatique IA, notamment la régression Lasso et divers modèles de machine learning (régression logistique, forêts aléatoires, réseaux de neurones, etc.) Les résultats sont satisfaisant et montrent que le modèle proposé est plus fiable que celui de Campusen. Pour avoir un model plus performant et utilisable les chercheurs envisagent de collecter plus de données sur les étudiants, notamment des données sur leur parcours académique antérieurs , intégrer des variables socio-économiques et psychologiques et enfin Finaliser et déployer une plateforme d'orientation intelligente pour l'enseignement supérieur sénégalais. Malheureusement je ne dispose je ne dispose pas des données nécessaire pour reproduire cette etude